HHS Public Access

Author manuscript

Graphs Biomed Image Anal Comput Anat Imaging Genet (2017). Author manuscript; available in PMC 2018 October 04.

Published in final edited form as:

Graphs Biomed Image Anal Comput Anat Imaging Genet (2017). 2017 ; 10551: 220–229. doi: 10.1007/978-3-319-67675-3_20.

Transcriptome-Guided Imaging Genetic Analysis via a Novel Sparse CCA Algorithm

Kefei Liu[1], **Xiaohui Yao**[1,2], **Jingwen Yan**[1,2], **Danai Chasioti**[1,2], **Shannon Risacher**[1], **Kwangsik Nho**[1], **Andrew Saykin**[1], **Li Shen**[1,2,*], and **for the Alzheimer's Disease Neuroimaging Initiative**[**]

[1]Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN 46202, USA

[2]School of Informatics and Computing, Indiana University, Indianapolis, IN, USA

Abstract

Imaging genetics is an emerging field that studies the influence of genetic variation on brain structure and function. The major task is to examine the association between genetic markers such as single nucleotide polymorphisms (SNPs) and quantitative traits (QTs) extracted from neuroimaging data. Sparse canonical correlation analysis (SCCA) is a bi-multivariate technique used in imaging genetics to identify complex multi-SNP-multi-QT associations. In imaging genetics, genes associated with a phenotype should at least expressed in the phenotypical region. We study the association between the genotype and amyloid imaging data and propose a transcriptome-guided SCCA framework that incorporates the gene expression information into the SCCA criterion. An alternating optimization method is used to solve the formulated problem. Although the problem is not biconcave, a closed-form solution has been found for each subproblem. The results on real data show that using the gene expression data to guide the feature selection facilities the detection of genetic markers that are not only associated with the identified QTs, but also highly expressed there.

1 Introduction

Brain imaging genetics is an emerging research field that studies the influence of genetic variation on brain structure and function. A fundamental problem in brain imaging genetics is to investigate the association between genetic variations such as single nucleotide polymorphisms (SNPs) and phenotypes extracted from multimodal neuroimaging data (e.g., anatomical, functional and molecular imaging scans). Given the well-known importance of gene and imaging phenotype in brain function, bridging these two factors and exploring their

*Correspondence to Li Shen (shenli@iu.edu).

**Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

connections would lead to a better mechanistic understanding of normal or disordered brain functions.

Sparse canonical correlation analysis (SCCA) has been widely adopted to identify complex imaging genetic associations in both synthetic and real imaging genetics data [2, 3, 6–8]. The prior graph or group structural knowledge among variables (e.g., a number of genes form a group to participate in a particular biological process to perform certain functionality in a cell) can be incorporated into SCCA model to guide the association analysis [2, 3, 9], which can improve the accuracy and stability in variable selection and facilitate the interpretability of the identified associations.

In this work, we propose to take advantage of the brain wide gene expression profile available in Allen human brain atlas (AHBA) and use it as a 2-D prior to guide the brain imaging genetics association analysis. To account for such 2-D prior, we propose a transcriptome-guided SCCA (TG-SCCA) framework that incorporates the gene expression information into traditional SCCA model. A new regularization term is introduced to encourage the discovery of imaging genomic associations so that the identified genes have relatively high expression level in their associated brain regions. To solve the formulated problem, we employ an alternating optimization method and manage to find a closed-form globally maximum solution for each of the two subproblems despite not biconcave.

Notation

The superscript $^{\mathrm{T}}$ stands for the transpose of a matrix or vector. The $\|\boldsymbol{u}\|$ and $\|\boldsymbol{u}\|_1$ denote the Euclidean norm and ℓ_1 norm of a vector $\boldsymbol{u}$, respectively. The operator $\odot$ represents the Hadamard product (entrywise product) of two matrices of the same dimensions. The sign function of a real number x is defined as follows: sign $\{x\} = 1$ when $x \geq 0$ and sign $\{x\} = -1$ when $x < 0$.

2 Problem Formulation

Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be the genotype data (SNP) and $\boldsymbol{Y} \in \mathbb{R}^{n \times q}$ be the imaging quantitative traits (QT) data, where n, p and q are the numbers of participants, SNPs and QTs, respectively. Sparse canonical correlation analysis (SCCA) aims to find a linear combination of variables in $\boldsymbol{X}$ and $\boldsymbol{Y}$ to maximize the correlation:

$$\underset{\boldsymbol{u},\boldsymbol{v}}{\text{maximize}}\, \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v} \quad \text{subject to} \quad \|\boldsymbol{u}\|^2 \leq 1, \|\boldsymbol{v}\|^2 \leq 1, \|\boldsymbol{u}\|_1 \leq c_1, \|\boldsymbol{v}\|_1 \leq c_2\,. \quad (1)$$

Suppose the q SNPs belong to G genes. Let $\boldsymbol{E} = \{e_{gj}\} \in \mathbb{R}^{G \times q}$ be the gene expression matrix with e_{gj} being the expression of gene g in brain region j. Let the SNPs be ordered in terms of the genes they belong to. Denote $\boldsymbol{u} = [\boldsymbol{u}_1^{\mathrm{T}}\boldsymbol{u}_2^{\mathrm{T}}\cdots\boldsymbol{u}_G^{\mathrm{T}}]^{\mathrm{T}}$, where $\boldsymbol{u}_g = [u_{g,1}\ u_{g,2}\ \cdots\ u_{g,p_g}]^{\mathrm{T}} \in \mathbb{R}^{p_g \times 1}$, $g = 1, 2, \ldots, G$, contains the canonical weights of the p_g SNPs in gene g, where p_g is the number of SNPs in gene g. To exploit the gene expression information, we propose to extend the SCCA in (1) in the following way:

$$\underset{\boldsymbol{u},\boldsymbol{v}}{\text{maximize}} \quad (1-\lambda)\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v} + \lambda \sum_{g=1}^{G}\sum_{j=1}^{q} \max\{|u_{g,1}|, |u_{g,2}|, \ldots, |u_{g,p_g}|\} e_{gj}|v_j| \quad (2)$$

$$\text{subject to} \quad \|\boldsymbol{u}\|^2 \le 1, \|\boldsymbol{v}\|^2 \le 1, \|\boldsymbol{u}\|_1 \le c_1, \|\boldsymbol{v}\|_1 \le c_2,$$

where $\boldsymbol{E} = \{e_{gj}\}$ is the gene expression matrix in which all the elements are equal to or greater than zero, and $\lambda \in [0, 1)$ is the weighting coefficient that reflects our confidence in imposing the correlation with the gene expression data. The regularization term encourages the selection of one SNP from each gene with relatively high expression in the relevant QTs. The intuition is that if a gene is related to a QT, it should be expressed in the corresponding brain tissue.

3 Methods

We employ an alternating optimization method to solve problem (2). Alternating optimization is an iterative procedure that proceeds in two alternating steps: update of $\boldsymbol{u}$ while holding $\boldsymbol{v}$ fixed and update of $\boldsymbol{v}$ while holding $\boldsymbol{u}$ fixed.

3.1 Update of u with v fixed

Denote

$$\boldsymbol{a} = (1-\lambda)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v} \in \mathbb{R}^{p\times 1}, \quad \boldsymbol{b} = \lambda \boldsymbol{E}|\boldsymbol{v}| \in \mathbb{R}^{G\times 1}.$$

The optimization of (2) with respect to $\boldsymbol{u}$ can be written as

$$\underset{\boldsymbol{u}}{\text{maximize}}\, \boldsymbol{a}^{\mathrm{T}}\boldsymbol{u} + \sum_{g=1}^{G} b_g \max\{|u_{g,1}|, |u_{g,2}|, \ldots, |u_{g,p_g}|\} \quad \text{subject to } \|\boldsymbol{u}\|^2 \le 1, \|\boldsymbol{u}\|_1 \le c_1.$$

(3)

The problem in (3) is highly non-concave. Notice however, that if we know which is largest among the absolute values of the optimal $u_{g,1}, u_{g,2}, \ldots, u_{g,p_g}$, we can narrow down our search space for optimal solutions and move $|u_{g,1}|, |u_{g,2}|, \ldots, |u_{g,p_g}|$ outside the max {} operator in the objective function. Next we determine the largest absolute value among the optimal $u_{g,1}, u_{g,2}, \ldots, u_{g,p_g}$.

Define

$$h(\boldsymbol{u}_g) = \boldsymbol{a}_g^{\mathrm{T}}\boldsymbol{u}_g + b_g \max\{|u_{g,1}|, |u_{g,2}|, \ldots, |u_{g,p_g}|\}, \quad (4)$$

where $\boldsymbol{a}_g \in \mathbb{R}^{p_g \times 1}$ is the subvector of $\boldsymbol{a}$ corresponding to the SNPs in gene g. The objective function in (3) can be written as $\sum_{g=1}^{G} h(\boldsymbol{u}_g)$.

Consider the optimization problem:

$$\underset{\boldsymbol{u}_g}{\text{maximize}}\, h(\boldsymbol{u}_g) \quad \text{subject to } \|\boldsymbol{u}_g\|^2 \le \mu^2, \|\boldsymbol{u}_g\|_1 \le \nu, \quad (5)$$

where μ and ν are arbitrary positive constants.

Without loss of generality suppose

$$|a_{g,1}| = \max\{|a_{g,1}|, |a_{g,2}|, \ldots, |a_{g,p_g}|\}.$$

It can be shown that the optimal solutions of problem (5) satisfy

$$|u_{g,1}| = \max\{|u_{g,1}|, |u_{g,2}|, \ldots, |u_{g,p_g}|\}. \quad (6)$$

Equation (6) can be proved by contradiction. The idea is that if equation (6) does not hold, the objective value $h(\boldsymbol{u}_g)$ can be increased by swapping the absolute values of the first element and the element with the largest magnitude in $\boldsymbol{u}_g$. Moreover, $u_{g,1}$ has the same sign as $a_{g,1}$; otherwise, reversing the sign of $u_{g,1}$ increases the objective value, which contradicts $\boldsymbol{u}_g$ being the optimal solution.

Therefore, the objective function for problem (5) becomes

$$h(\boldsymbol{u}_g) = (a_{g,1} + b_g \operatorname{sign}\{a_{g,1}\})u_{g,1} + a_{g,2}u_{g,2} + \cdots + a_{g,p_g}u_{g,p_g}.$$

For $g = 1, 2, \ldots, G$, let

$$\ell_g = \underset{k=1,2,\ldots,p_g}{\arg\max} |a_{g,k}| \quad (7)$$

and let $\boldsymbol{e}_{\ell_g}$ be a length-p_g column vector with 1 at location ℓ_g and 0 elsewhere.

The problem (3) reduces to solving

$$\underset{\boldsymbol{u}}{\text{maximize}} \sum_{g=1}^{G} [\boldsymbol{a}_g + b_g \boldsymbol{e}_{\ell_g} \odot \operatorname{sign}\{\boldsymbol{a}_g\}]^{\mathrm{T}} \boldsymbol{u}_g \quad \text{subject to } \|\boldsymbol{u}\|^2 \le 1, \|\boldsymbol{u}\|_1 \le c_1. \quad (8)$$

Define

$$\boldsymbol{w} = \begin{pmatrix} \boldsymbol{a}_1 + b_1 \boldsymbol{e}_{\ell_1} \odot \text{sign}\{\boldsymbol{a}_1\} \\ \boldsymbol{a}_2 + b_2 \boldsymbol{e}_{\ell_2} \odot \text{sign}\{\boldsymbol{a}_2\} \\ \vdots \\ \boldsymbol{a}_G + b_G \boldsymbol{e}_{\ell_G} \odot \text{sign}\{\boldsymbol{a}_G\} \end{pmatrix}. \quad (9)$$

The problem (8) is expressed in a more compact form as

$$\underset{\boldsymbol{u}}{\text{maximize}}\, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{u} \quad \text{subject to } \|\boldsymbol{u}\|^2 \le 1, \|\boldsymbol{u}\|_1 \le c_1. \quad (10)$$

According to [7, Lemma 2.2], problem (10) has a closed-form solution which can be obtained by shrinking the elements in $\boldsymbol{w}$ toward zero by a non-negative constant and then normalizing the result to unit norm. Formally, the solution to (8) is

$$\boldsymbol{u}^* = \frac{S(\boldsymbol{w}, \Delta)}{\|S(\boldsymbol{w}, \Delta)\|} \quad (11)$$

with $\Delta = 0$ if this results in $\|\boldsymbol{u}^*\|_1 \le c_1$; otherwise, Δ is a positive number that satisfies $\|\boldsymbol{u}^*\|_1 = c_1$. In (11), $S(\boldsymbol{w}, \Delta)$ is the soft-thresholding operator that is applied to each element of $\boldsymbol{w}$, with

$$S(w_i, \Delta) = \begin{cases} w_i - \Delta, & w_i > \Delta \\ w_i + \Delta, & w_i < -\Delta \\ 0, & -\Delta \le w_i \le \Delta \end{cases}$$

3.2 Update of $\boldsymbol{v}$ with $\boldsymbol{u}$ fixed

Denote

$$\boldsymbol{\alpha} = (1-\lambda)\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u} \in \mathbb{R}^{q \times 1}.$$

The optimization of (2) with respect to $\boldsymbol{v}$ can be written as

$$\underset{\boldsymbol{v}}{\text{maximize}} \sum_{j=1}^{q} \alpha_j v_j + \beta_j |v_j| \quad \text{subject to } \|\boldsymbol{v}\|^2 \le 1, \|\boldsymbol{v}\|_1 \le c_2. \quad (12)$$

where $\beta_j = \lambda \sum_{g=1}^{G} \max\{|u_{g,1}|, |u_{g,2}|, \ldots, |u_{g,p_g}|\} e_{gj}$, for $j = 1, 2, \ldots, q$.

Since $\beta_j \geq 0$ it can be shown that the optimal v_j has the same sign as α_j[2].

Define

$$\boldsymbol{\gamma} = \begin{pmatrix} \alpha_1 + \beta_1 \operatorname{sign}\{\alpha_1\} \\ \alpha_2 + \beta_2 \operatorname{sign}\{\alpha_2\} \\ \vdots \\ \alpha_q + \beta_q \operatorname{sign}\{\alpha_q\} \end{pmatrix}. \quad (13)$$

The optimization problem in (12) boils down to solving the following problem:

$$\underset{\boldsymbol{v}}{\text{maximize}}\, \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{v} \quad \text{subject to } \|\boldsymbol{v}\|^2 \leq 1, \|\boldsymbol{v}\|_1 \leq c_2. \quad (14)$$

According to [7, Lemma 2.2], the solution to (14) is

$$\boldsymbol{v}^* = \frac{S(\boldsymbol{\gamma}, \delta)}{\|S(\boldsymbol{\gamma}, \delta)\|} \quad (15)$$

with $\delta = 0$ if this results in $\|\boldsymbol{v}^*\|_1 \leq c_2$; otherwise, δ is a positive number that satisfies $\|\boldsymbol{v}^*\|_1 = c_2$.

Given an initial estimate for $\boldsymbol{v}$, the TG-SCCA algorithm alternately update $\boldsymbol{u}$ and $\boldsymbol{v}$ in an iterative manner until convergence, as outlined in Algorithm 1.

Remark 1—Analysis shows that the optimization problem (2) is not biconcave in **u** and **v**: it is neither concave in **u** when **v** is fixed, nor concave in **v** when **u** is fixed[3]. Interestingly, the TG-SCCA algorithm finds the global maxima of the two subproblems in each iteration (i.e., Steps 5 and 6 of Algorithm 1).

Algorithm 1

TG-SCCA algorithm

Input: Genotype data: $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, imaging phenotype data $\boldsymbol{Y} \in \mathbb{R}^{n \times q}$, and gene expression data $\boldsymbol{E} \in \mathbb{R}^{G \times q}$;

1: Normalize the columns of $\boldsymbol{X}$ and $\boldsymbol{Y}$ to have zero mean and unit Euclidian norm;

2: Choose the tuning parameters c_1, c_2 and λ;

3: Initialize $\boldsymbol{u} \in \mathbb{R}^{p \times 1}$ and $\boldsymbol{v} \in \mathbb{R}^{q \times 1}$;

4: **repeat**

5: Update $\boldsymbol{u}$ according to Eqs. (9) and (11);

[2]Otherwise, we can always increase the objective value by reversing the sign of v_i.
[3]The problem (2) is actually biconvex in **u** and **v**.

6: Update $\boldsymbol{v}$ according to Eqs. (13) and (15);

7: **until** convergence.

4 Experimental results and discussions

We compare the TG-SCCA algorithm with the conventional SCCA algorithm [7] on a real imaging genetics data set to demonstrate its performance. The genotyping and baseline AV-45 PET data of 774 non-Hispanic Caucasian subjects, including 187 healthy control (HC), 76 significant memory concern (SMC), 227 early mild cognitive impairment (MCI), 186 late MCI, and 98 AD participants, were downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. One aim of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, see www.adni-info.org.

The AV-45 images were aligned to each participant's same visit MRI scan and normalized to the Montreal Neurological Institute (MNI) space. Region-of-interest (ROI) level AV-45 measurements were further extracted based on the MarsBaR AAL atlas. In this study, we focused on the analysis of 1,221 SNPs from 56 AD risk genes and 78 AD related ROIs. Using the regression weights derived from the HC participants, the genotype and imaging measures were preadjusted for removing the effects of the baseline age, gender, education, and handedness. The gene expression data were obtained from the Allen Human Brain Atlas (human.brain-map.org). The data were already normalized via a series of processes to remove non-biological biases and to make those comparable across samples. See Fig. 1 for the expression levels of the studied genes in the studied ROIs.

4.1 Selection of tuning parameters λ, c_1 and c_2

Based on our observation that the ℓ_1-sparsity constraints in the SCCA model (1) are not active when $c_1 \geq \|\boldsymbol{u}_1\|_1$ and $c_2 \geq \|\boldsymbol{v}_1\|_1$, where $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ are the left and right singular vectors of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ corresponding to its largest singular value, we propose to set the parameters c_1 and c_2 in the following way: $c_1 = s_1 \|\boldsymbol{u}_1\|_1$ and $c_2 = s_2 \|\boldsymbol{v}_1\|_1$, where $0 < s_1, s_2 < 1$. For the TG-SCCA, we set $c_1 = s_1 [(1 - \lambda) \|\boldsymbol{u}_1\|_1 + \lambda \|\boldsymbol{u}_2\|_1]$ and $c_2 = s_2 [(1 - \lambda) \|\boldsymbol{v}_1\|_1 + \lambda \|\boldsymbol{v}_2\|_1]$, where $\boldsymbol{u}_2$ and $\boldsymbol{v}_2$ are the left and right singular vectors of $\boldsymbol{E}$ corresponding to its largest singular value and $\boldsymbol{E}$ has been scaled to have the same spectral norm as that of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$. We set $s_1 = s_2 = 0.5$, and $\lambda = 0.5$. Note that the sparsity level should not affect the relative performance of the SCCA and TG-SCCA algorithms as long as the same sparsity is used for them.

To improve the robustness to the particular choice of the tuning parameters and variable selection accuracy, we employ stability selection [5], which fits the SCCA model to a large number of (100) random subsamples, each of size $n/2$. Variable selection results across all subsamples are integrated to compute the empirical selection probability for each genetic and imaging variable.

4.2 Results

Fig. 2 shows the empirical selection probability of the top 25 SNPs and top 10 QTs of the SCCA and TG-SCCA algorithms applied to the AV45 data, and the map of expression profile of the identified genes in the identified brain regions. Six genes (*CLU, CST3, MEF2C, PRNP, SORL1* and *THRA*) are detected by TG-SCCA but not by SCCA. For most of these genes, evidence has been reported in the literature on their association with AV-45 measures. For example, *CST3* risk haplotype may account for greater amyloid load or neuronal death and affect resting cortical rhythmicity [1], and the association of *SORL1* gene with hippocampal and cerebral atrophy was reported in [4].

Fig. 3 shows the histograms of bootstrapped correlation coefficients from the analysis of the AV45 data, indicating the correlation strength detected by TG-SCCA is similar to that by SCCA. While maintaining a similar correlation discovery power (Fig. 3), TG-SCCA identifies a set of imaging and genetic markers so that the average expression level of the identified genes in the identified regions is much higher than that obtained in SCCA (Fig. 2(c)). This meets our expectation. Given these high expression profiles, the identified imaging genetic associations have the potential to provide improved mechanistic understanding of genetic basis of AD-related brain imaging phenotypes.

5 Conclusions

Many existing studies first identify the imaging genetic associations and then go to Allen Human Brain Atlas (AHBA) to look for additional evidence (i.e., the identified gene is expressed in the relevant region). In this work, we have coupled these two steps together and propose a transcriptome-guided sparse canonical correlation analysis (TG-SCCA) framework that directly identifies strong imaging genetic associations with transcriptomic support evidenced in AHBA. To solve the formulated problem, we have developed an efficient algorithm which finds a closed-form global solution for each of the two subproblems. Our study on real imaging genetics data in an AD study has demonstrated that TG-SCCA yields promising and biologically meaningful findings.

Acknowledgments

This work was supported by NIH R01 EB022574, R01 LM011360, U01 AG024904, P30 AG10133, R01 AG19771, UL1 TR001108, R01 AG 042437, R01 AG046171, and R01 AG040770, by DoD W81XWH-14-2-0151, W81XWH-13-1-0259, W81XWH-12-2-0012, and NCAA 14132004.

References

1. Braskie MN, Ringman JM, Thompson PM. Neuroimaging measures as endophenotypes in Alzheimer's disease. International journal of Alzheimer's disease. Mar.2011 2011:490140.
2. Chen J, Bushman FD, Lewis JD, Wu GD, Li H. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. Biostatistics. 2013; 14(2):244–258. [PubMed: 23074263]
3. Chen X, Liu H, Carbonell JG. International Conference on Artificial Intelligence and Statistics. Vol. 12. La Palma, Canary Islands: 2012. Structured sparse canonical correlation analysis; 199–207.
4. Louwersheimer E, Ramirez A, Cruchaga C, Becker T, Kornhuber J, Peters O, Heilmann S, Wiltfang J, Jessen F, Visser PJ, Scheltens P, Pijnenburg YAL, Teunissen CE, Barkhof F, van Swieten JC,

Holstege H, Van der Flier WM. A. D. N. Initiative, and D. C. Network. Influence of genetic variants in SORL1 gene on the manifestation of Alzheimer's disease. Neurobiol. Aging. Mar.2015 36:1605.e13–1605.e20.

5. Meinshausen N, Bühlmann P. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2010; 72(4):417–473.
6. Parkhomenko E, Tritchler D, Beyene J. Sparse canonical correlation analysis with application to genomic data integration. Stat. Appl. Genet. Mol. Biol. 2009; 8:1–34.
7. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. 2009; 10(3):515–34. [PubMed: 19377034]
8. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. Stat. Appl. Genet. Mol. Biol. 2009; 8(1):1–27.
9. Yan J, Du L, Kim S, Risacher SL, Huang H, Moore JH, Saykin AJ, Shen L. Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. Bioinformatics. 2014; 30(17):i564–i571. [PubMed: 25161248]

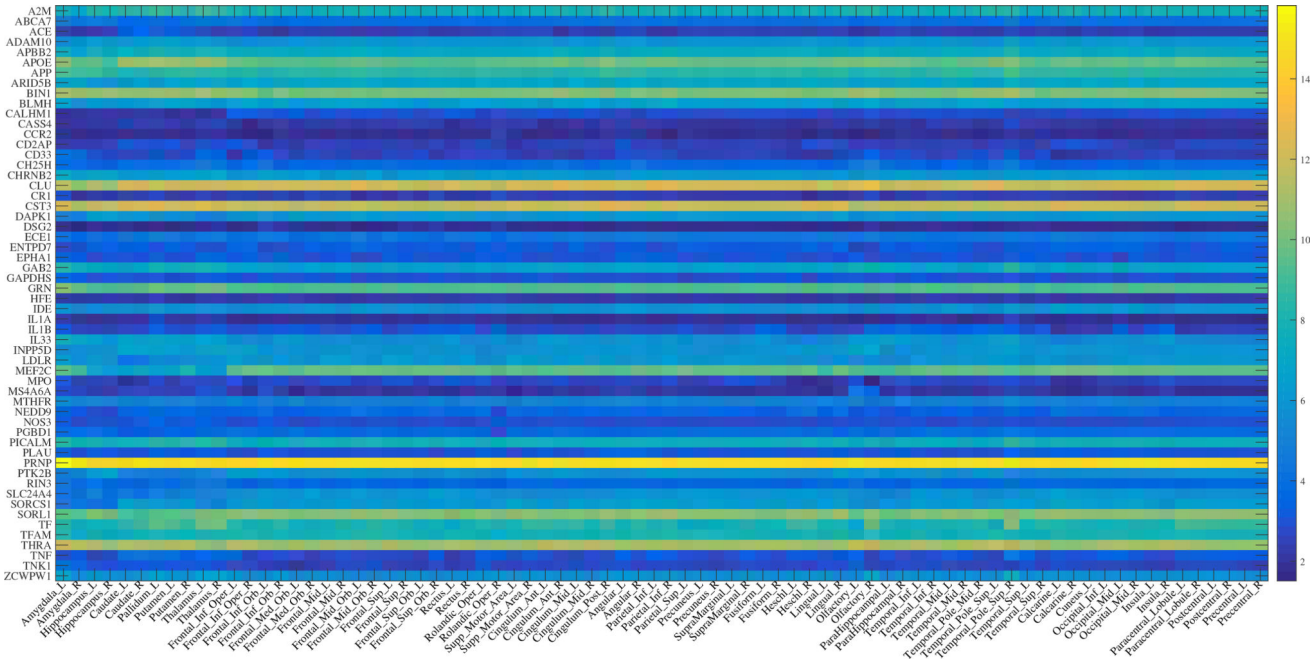

Fig. 1.

Brain transcriptome: Map of expression level of the studied genes in the brain regions of interest.

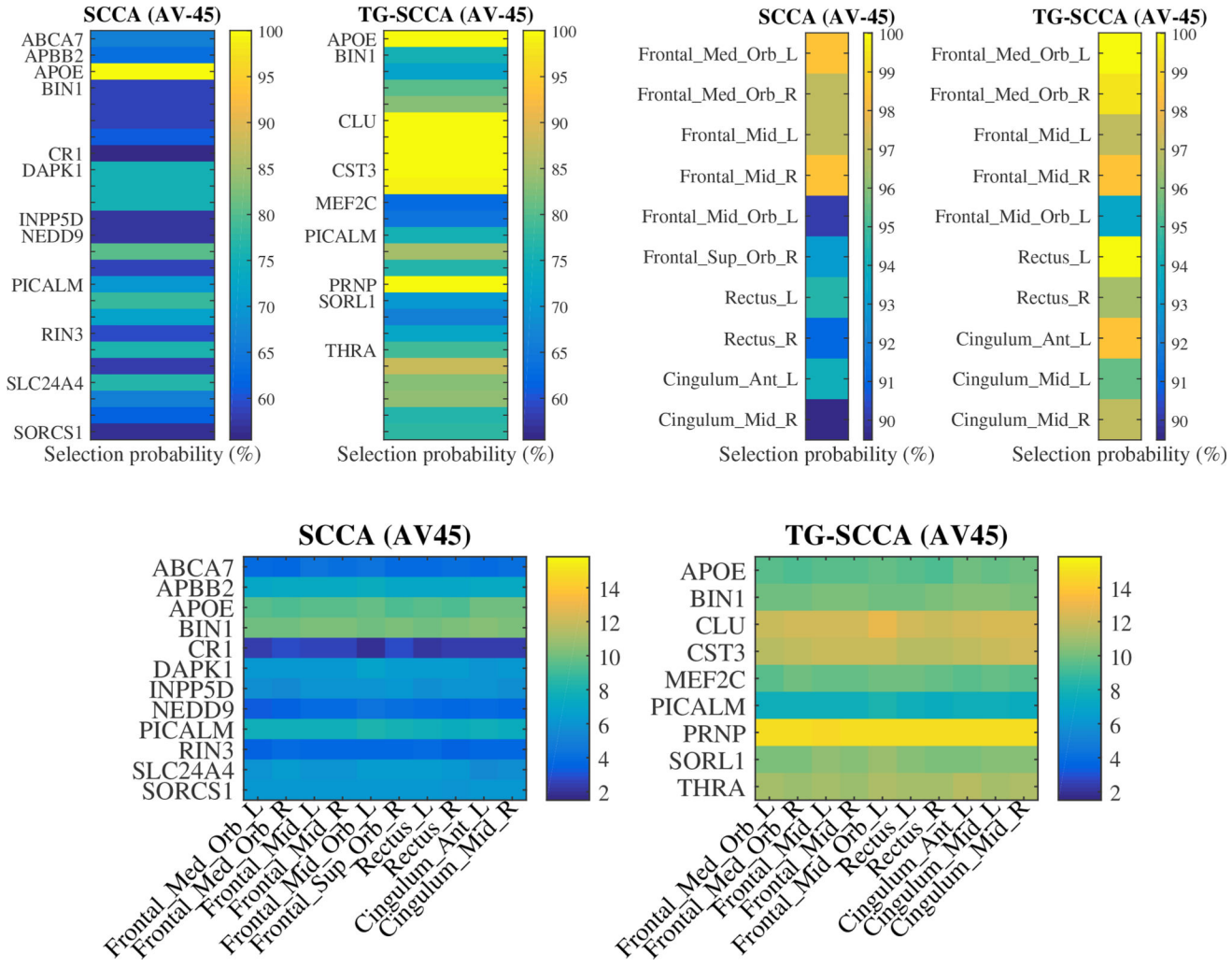

Fig. 2.

Comparison of feature selection results between SCCA and TG-SCCA: (a) Selection probability map of top 25 identified SNPs, labelled with their corresponding genes. Each SNP belongs to the nearest gene above it on the heatmap. (b) Selection probability map of top 10 identified imaging biomarkers. (c) Expression level of the identified genes in the identified ROIs.

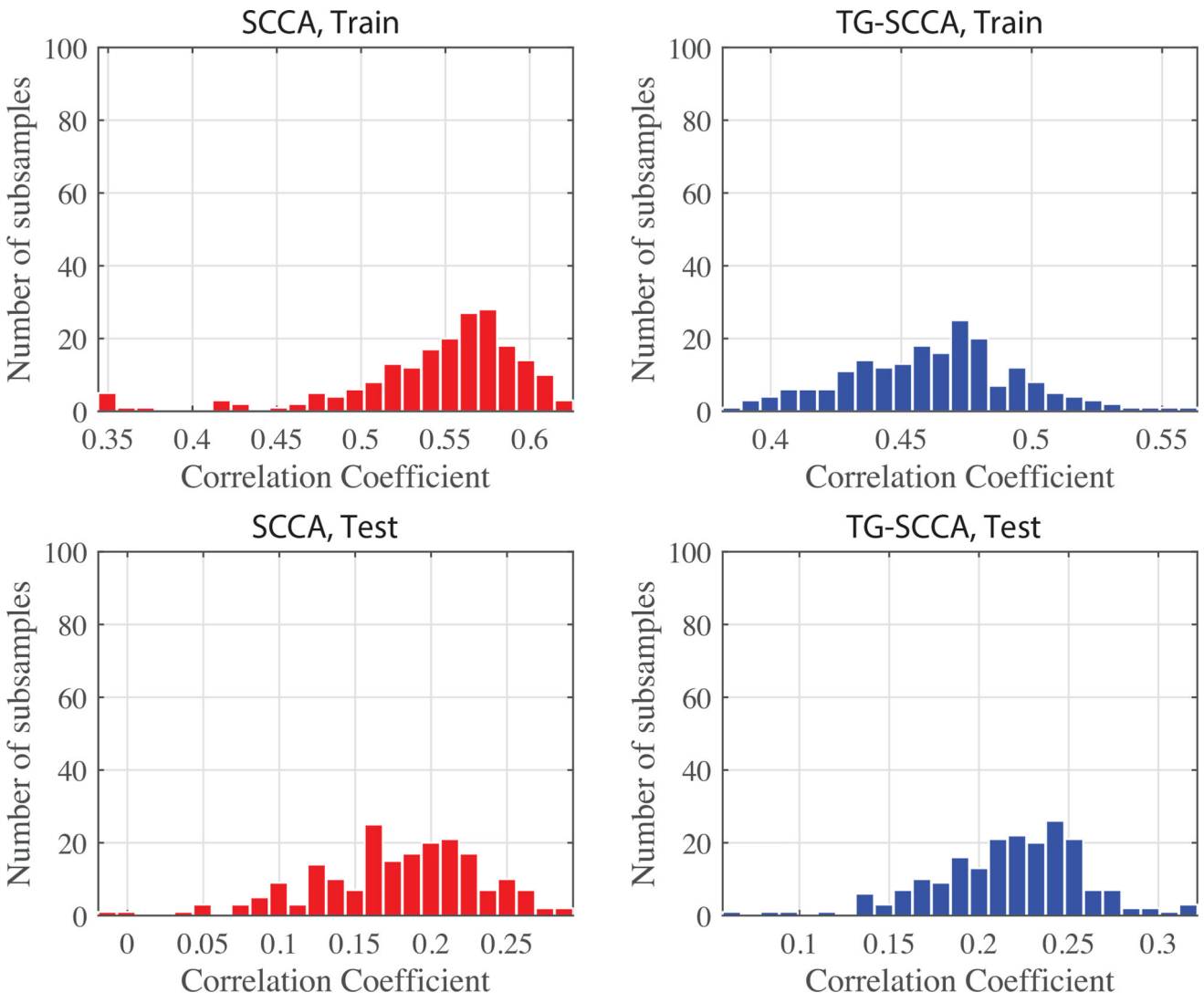

Fig. 3.

Histograms of correlation coefficients from analysis of AV45 data, with training (left) and test (right) results of the SCCA (left) and TG-SCCA (right) being shown.